DOCUMENT RESUME

ED 213 731 TM 820 075

AUTHOR Wilcox, Rand R.
TITLE An Alternative Interpretation of Three Stability Models. Measurement and Methodology, Work Unit 2: Technical Adequacy of Tests.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Jun 78
NOTE 11p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Comparative Analysis; *Error of Measurement; *Mathematical Models; *Probability; Statistical Analysis; Testing Problems; *Test Items; Test Theory; *True Scores

ABSTRACT

Two fundamental problems in mental test theory are to estimate true score and to estimate the amount of error when testing an examinee. In this report, three probability models which characterize a single test item in terms of a population of examinees are described. How these models may be modified to characterize a single examinee in terms of an item domain is explained. A numerical illustration is given which shows how the modified models might be used to adjust observed scores when comparing examinees. (Author/AL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED213731

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

DELIVERABLE: June 1978

MEASUREMENT AND METHODOLOGY

WORK UNIT 2

Technical Adequacy of Tests

Rand R. Wilcox
Project Director

AN ALTERNATIVE INTERPRETATION OF THREE STABILITY MODELS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. S. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 075

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California - Los Angeles

June 1978

"The activity which is the subject of this report was supported in part or in whole by the National Institute of Education, Department of HEW. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred."

ABSTRACT

Wilcox (1977) describes three probability models which characterize a single test item in terms of a population of examinees. The purpose of this note is to indicate that similar models can be derived which characterize a single examinee in terms of an item domain. A numerical illustration is given which shows how our models might be used to adjust observed scores when comparing examinees.

CSE
ARCHIVE COPY

When describing an examinee in terms of a single achievement test item, it may be reasonable in certain situations to say that an examinee either "knows" or "does not know" the correct response. One method of deriving this two-class-latent-structure model is to assume that learning is all or none (Macready and Dayton, 1977). However, alternative derivations are certainly possible. Suppose, for example, that learning can be characterized by a continuous latent variable θ, where $-\infty < \theta < \infty$. As θ gets large, the probability of a correct response goes up. One possibility is to define "knowing" as having $\theta > \theta_0$ where θ_0 is some appropriately defined constant. We are not suggesting that this be done or that this is a desirable method of defining what it means to know the answer to a particular achievement test item. We are merely pointing out that we are not necessarily taking an all-or-none view of learning when we say that an examinee either knows or does not the correct response.

When measuring achievement it is often desired to describe an examinee in terms of an item domain (see, e.g., Harris, 1974). When for each item in the item domain it is meaningful to say that an examinee either knows or does not know the correct response, a natural definition of true score is the proportion of items the examinee knows. The difficulty is that an examinee's observed responses might not reflect his true state. That is, the examinee may know the correct answer but due to some distraction give an incorrect response. Alternatively, the examinee may guess the correct response when in fact he does not know. Two fundamental problems in mental test theory are to estimate true score and to estimate the amount of error when testing an examinee. We describe how three probability models given by Wilcox (1977) may be modified to solve these problems for the situation described above.

Let k represent the proportion of items in the item domain which the examinee knows. The item pool might consist, for example, of the problem of solving equations of the form ax+b=cx+d. Hively et al. (1973) describes a method which might be used to randomly generate the test items. For this case, k represents the proportion of all such items for which the examinee knows the appropriate skills to arrive at the correct response. We further assume that for a randomly selected item, the examinee guesses the correct answer to an item he does not know with probability y. For an item the examinee knows, x represents the probability of an incorrect response.

Suppose the examinee responds to the same randomly chosen item on two separate occasions in time. In our example, we are considering a randomly generated equation of the form ax+b=cx+d. Also assume that there is no forgetting or learning between the two administrations of the test and that the results on the first administration are independent of the results on the second. For a randomly chosen item, let

p_{11} = the probability of a correct response on both occasions

p_{10} = the probability of a correct-incorrect response

p_{01} = the probability of an incorrect-correct response

p_{00} = the probability of an incorrect response on both occasions.

Based on the assumptions made,

$$p_{11} = k(1-x)^2 + (1-k)y^2$$

$$(1) \qquad p_{10} = p_{01} = k(1-x)x + (1-k)y(1-y)$$

$$p_{00} = kx^2 + (1-k)(1-y)^2$$

As indicated, $p_{10} = p_{01}$. This requirement of our probability model can be tested with McNemar's test (McNemar, 1947). Wilcox (1977) gives some results on the power function of McNemar's test which may be useful when testing this hypothesis.

If it were possible to express k, x and y as a function of p_{11}, p_{10}, p_{01}, p_{00} it would be possible to estimate k, x and y based on the observed responses of the examinee. This is impossible though since we have in effect two equations in three unknowns. However, by imposing restrictions on the parameters x and y, solutions are possible. One possible assumption is x=0, i.e., if the examinee knows, he is always correct. In this case we can solve (1) for k and y which gives

$$y = p(p+p_{00})^{-1}$$

$$k = 1-(p+p_{00})^{2}\, p_{00}^{-1},$$

where $p = p_{10} = p_{01}$.

A solution is also possible by assuming y=0, x > 0. The assumption that it is impossible to not know and guess may be appropriate when completion items are used. Solving for x and k gives

$$x = p(p_{11} + p)^{-1}$$

$$k = p_{11}^{-1}\,(p_{11} + p)^{2}.$$

Still another analytic solution is possible by setting x=y which yields the Lazarsfeld-Kendall "turnover" model as described by Goodman and Kruskal (1959, p. 150). As noted by Goodman and Kruskal, the necessary and sufficient conditions for this model to hold are

$$p_{10} \leq \tfrac{1}{2}$$

$$p_{11} = p_{01}$$

$$p_{11} \geq (p_{11} + p_{10})(p_{11} + p_{01}).$$

Solving for x, y and k gives

$$x = y = \tfrac{1}{2}(1 \pm (1-4p)^{\frac{1}{2}})$$

$$k = \tfrac{1}{2} + \frac{2p_{1.} - 1}{2\sqrt{1-4p.}}$$

where $p_{1.} = p_{11} + p_{10}$.

We observe that the three models just described are similar to the misinformation model of tests (Duncan, 1974). In this model it is assumed that $y=a^{-1}$, where a is the number of alternatives for each multiple-choice item used. It is further assumed that the examinee has misinformation and hence is always incorrect on some fixed number of items on a particular test. The assumption $y=a^{-1}$ may be untenable in some situations since the examinee may be able to eliminate one or more of the alternatives from consideration without knowing the correct response. In our models this assumption is not made.

For the three probability models just described, the natural method of estimating x, y and k is to substitute the maximum likelihood estimates of p_{11}, p_{10}, p_{01} and p_{00} into the expressions for x, y and k. We observe that under certain circumstances, this method of estimation will yield maximum likelihood estimates of x, y and k. However, if we assume the probability models hold, this estimation procedure may yield an inadmissible value which means the estimate is not maximum likelihood. The results

given by Wilcox (1977) are readily applied to the present situation and so we do not discuss this point further.

Numerical Illustration

To illustrate how the models might be used as well as their possible implications, suppose we want to compare two examinees in terms of their achievement relative to a particular item domain. Suppose further that each examinee takes 50 randomly chosen items on two separate occasions resulting in the observed frequencies as shown in the two-way contingency tables.

EXAMINEE A

		Second Testing 1	Second Testing 0	
First Testing	1	29	6	35
	0	6	9	15
		35	15	50

EXAMINEE B

		Second Testing 1	Second Testing 0	
First Testing	1	30	1	31
	0	3	16	19
		33	17	50

We see, for example, that there were 3 items for which examinee B had an incorrect (0) and then a correct (1) response. Maximum likelihood estimates of p_{11}, p_{10}, p_{01}, p_{00} are $\hat{p}_{11}=29/50$, $\hat{p}_{10}=6/50$, $\hat{p}_{01}=6/50$ and $\hat{p}_{00}=9/50$. For examinee B they are $\hat{p}_{11}=30/50$, $\hat{p}_{10}=1/50$, $\hat{p}_{01}=3/50$, and $\hat{p}_{00}=16/50$. The assumption $p=p_{10}=p_{01}$ suggests that we estimate p with $\hat{p}=(p_{10}+p_{01})/2$. Wilcox (1977) shows that $\hat{p}$ is an efficient, maximum likelihood estimate of p when the model holds.

For the sake of illustration, suppose that because of the nature of the test items, $x = 0$. Assuming $x=0$ we may estimate k and y by substituting $\hat{p}$ and $\hat{p}_{00}$ into the appropriate expressions described above. For examinee A we estimate k and y to be $\hat{k}=.5$ and $\hat{y}=.4$. For examinee B we have $\hat{k}=.595$ and $\hat{y}=.111$. On both occasions the number-correct score of the first examinee is 35. For examinee B the test scores are 34 and 31. Thus, using the test results on either of the two test administrations, the number-correct scores suggest that examinee A should be ranked higher than examinee B. However, the estimates of k and y suggest that the reverse decision be made.

REFERENCES

Duncan, George T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

Goodman, L. A. and Kruskal, W. H. Measures of association for cross classifications II: Further discussion and references. Journal of the American Statistical Association, 1959, 54, 123-163.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin and W. James Popham (Eds.) Problems in criterion-referenced measurement. CSE Monograph No.3, Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hively, W., Maxwell, G., Rabehl, G., Sension, D. and Lundin, S. Domain - referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. CSE Monograph No.1, Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Macready, G. B. and Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

McNemar, Q. Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 1947, 12, 153-7.

Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. Pastorok and R. Wilcox, Achievement Test Items - Methods of Study. CSE Monograph No.6, Los Angeles: Center for the Study of Evaluation, University of California, 1977.